Weekly Report

1 Done

1.1 Review

- Steering Data Quality with Visual Analytics: the Complexity Challenge

This paper presents a visual analytics framework for steering data quality and summarizes challenges and opportunities for different data types. It makes an important and useful contribution. Data quality is a serious problem right now. The data is collected, but sometimes in a poor quality. I believe this paper can attract more researchers pay attention to this topic, which is of great significance.

Authors provides enough motivation. Actually, data sanitization always leads to low quality. However, it is necessary due to privacy preservation. Adding related contents and citations can make motivation more convincing.

Also, the entire paper is in a good organization. But it can be improved. Authors describe the features and potential problems that may lead to low quality in Section 4. My advice is extracting common issues rather than introducing repeatedly. For example, there exists common problems, like missing values, outliers and inaccurate values. I think that the structure in Section 6 is better. Or you can try to split it into two Sections for commons and differences respectively.

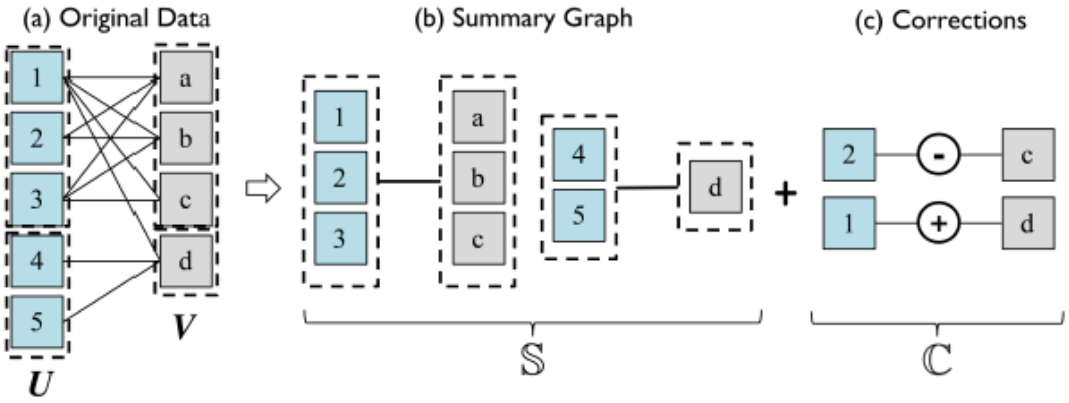
Besides, there are some typos need to be removed.

1. [5, 2] → [2, 5]
2. … and attributes (e.g. event magnitude, actors, etc.)[.]
3. et. al. → et al.
4. y , attributes → y, attributes
5. In Figure 1, all the words next to the arrow are verbs, except “projections”.
6. In Figure 2: wrong capitalization and etc[.].
7. Reference 2: the title of the journal should be capitalized.

1.2 Paper Reading

- VIBR: Visualizing Bipartite Relations at Scale with the Minimum Description Length Principle

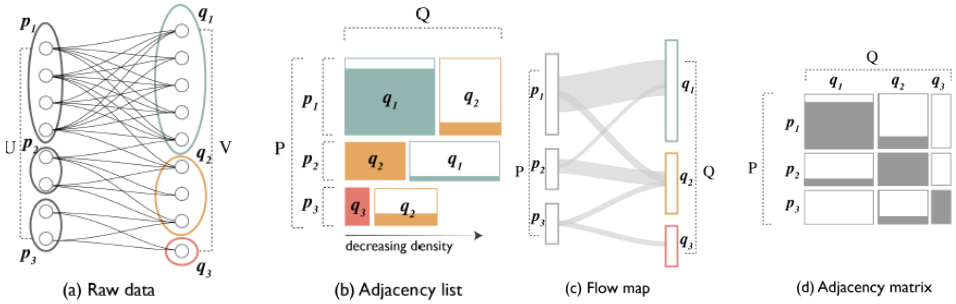
This work applies MDL (minimal description length) principle to pack large scale and noisy bipartite relation data.



So, the method seeks the minimal value of the L_R as shown below.

$$L_R(P,Q) = L(\mathbb{S}) + L(\mathbb{C})$$

Also, the results can be employed in different presentation.



I think the idea of this paper is similar to matrix compression. So, I wonder if this work can be extended to general graphs. Then, the compression can be used for both storage and representation.

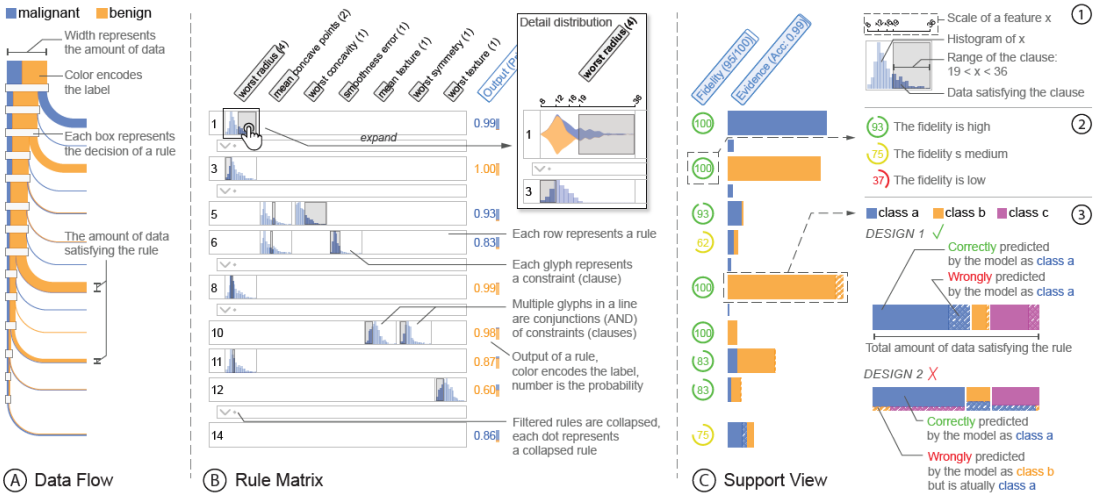
- RuleMatrix: Visualizing and Understanding Classifiers with Rules

This paper provides a visual technique that helps domain experts understand and inspect classification models using rule-based explanation. We need explain classification as well, so I read it to learn about the visual design.

Input: model F, training data $\mathcal{X}$, rule learning algorithm TRAIN
Parameters: parameter n_{sample}, feature set $\mathcal{S}$
Output: A rule list R that approximates F

1 $M \leftarrow$ ESTIMATEDISTRIBUTION($\mathcal{X}, \mathcal{S}$);
2 Draw samples $\mathcal{X}_{sample} \leftarrow$ SAMPLE($M, n_{samples}$);
3 Get the labels of $\mathcal{X}_{sample}$ using: $\mathcal{Y}_{sample} \leftarrow F(\mathcal{X}_{sample})$;
4 Rule list $R \leftarrow$ TRAIN($\mathcal{X}_{sample}, \mathcal{Y}_{sample}$);
5 **return** R;

Before that, for a trained classification model, they extract a rule list that approximates the original one using model induction. Based on those rules, they have design as shown as below.



A: The data flow visualizes the data that satisfies a rule as a flow into the rule, providing an overall sense of the order of the rules.

B: The rule matrix presents each rule as a row and each feature as a column. The clauses are visualized as glyphs in the corresponding cells. Users can click to expand a glyph to see the details of the distribution and the interval of the clause.

C: The support view shows the fidelity of the rule for the provided data, and the evidence of the model's predictions and errors under a certain rule.

2 Remarks

The mountain fire is about to go away, and the air has gotten better. But now, everyone is in a Thanksgiving holiday. Next week I will be able to go to the lab to register and start working.